NAAN MUDHALVAN -PHASE 1 PROJECT

PROJECT : Serverless IOT Data Processing.

AIM:

Design a serverless IoT data processing system on IBM Cloud for real-time scalability and automation, alongside exploring the role of Hadoop for batch processing and large data analytics.

PROBLEM STATEMENT:

what is this project all about?

Use serverless functions and an autonomous database in IBM Cloud to automate and scale the processing of streamed IoT data.

why we need this project?

Internet of Things (IoT) devices need to scale efficiently in real time.

By doing this in Real time as we deploy more devices and sensors, the volume and variety of streamed data is bound to grow.

PROBLEM DEFINITION AND DESIGN THINKING:

ARCHITECTURE:

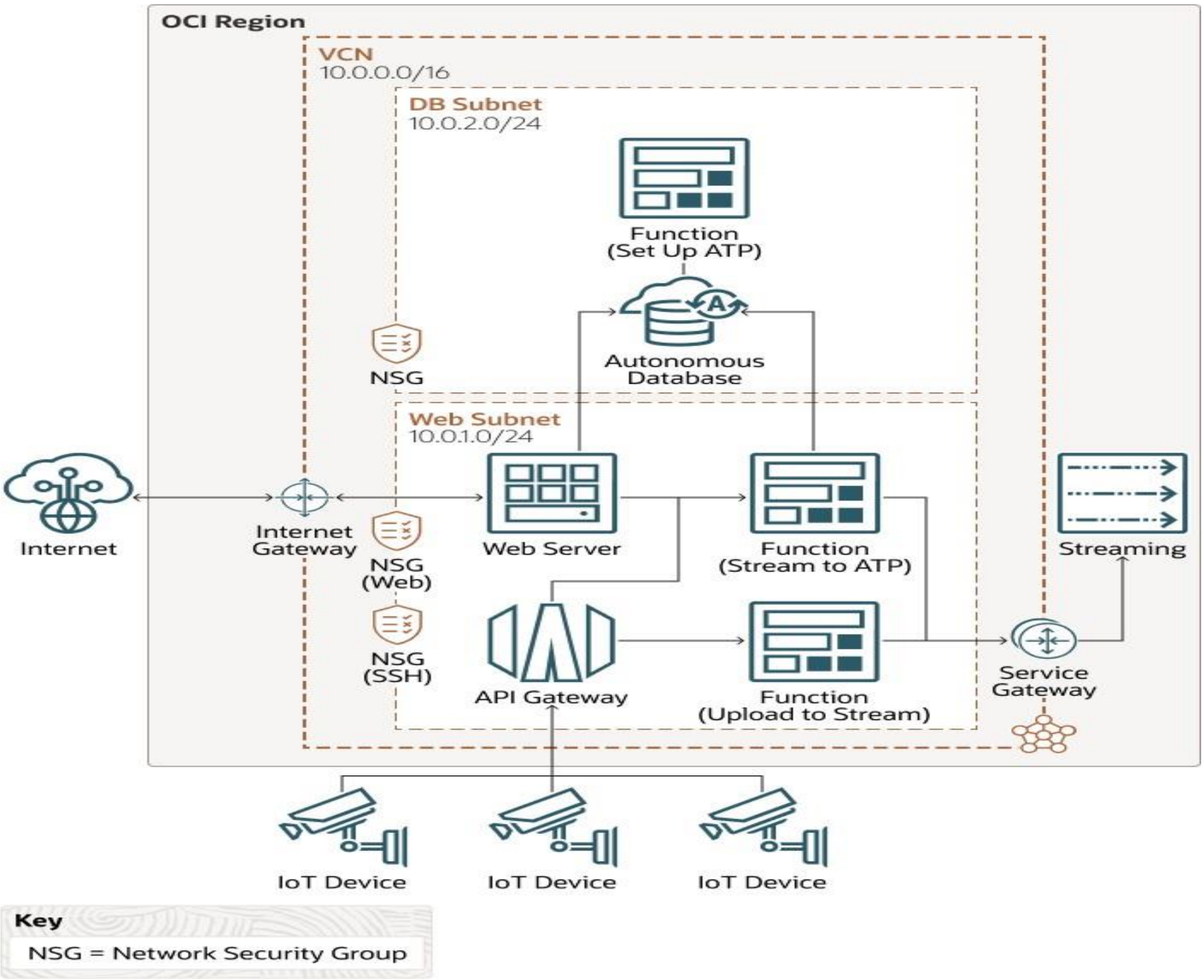
In this architecture, data from IoT devices flows in through an API gateway to serverless functions, which use the Streaming service to upload the data to an autonomous database in IBM Cloud. Users outside the cloud can access the data through a web server running on an IBM Cloud Infrastructure Compute instance.

- **Region:**

  An IBM Cloud Infrastructure region is a localized geographic area that contains one or more data centers, called availability domains. Regions are independent of other regions, and vast distances can separate them (across countries or even continents).

- **Virtual cloud network (VCN) and subnets:**
  A VCN is a customizable, software-defined network that you set up in an IBM Cloud Infrastructure region. Like traditional data center networks, VCNs give you complete control over your network

environment. A VCN can have multiple non-overlapping CIDR blocks that you can change after you create the VCN.It can segment a VCN into subnets, which can be scoped to a region.



- **Network security groups (NSG):**

NSGs act as virtual firewalls for your cloud resources. With the zero-trust security model of IBM Cloud Infrastructure, all traffic is denied, and you can control the network traffic inside a VCN. An NSG consists of a set of ingress and egress security rules that apply to only a specified set of VNICs in a single VCN.Access to the database and the web server in this architecture is controlled through separate NSGs.

- **API Gateway:**

IBM API Gateway enables you to publish APIs with private endpoints that are accessible from within your network, and which you can expose to the public internet if required. The endpoints support API validation, request and response transformation, CORS, authentication and authorization, and request limiting.

- **Streaming:**

IBM Cloud Infrastructure Streaming provides a fully managed, scalable, and durable storage solution for ingesting continuous, high-volume streams of data that you can consume and process in real time. You can use Streaming for ingesting high-volume data, such aapplication logs, operational telemetry, web click-stream data; or for other use cases where data is produced and processed continually and sequentially in a publish-subscribe messaging model.

- **Functions:**

IBM Functions is a fully managed, multitenant, highly scalable, on-demand, Functions-as-a-Service (FaaS) platform. It is powered by the Fn Project open source engine. Functions enable you to deploy your code, and either call it directly or trigger it in response to events. IBM Functions uses Docker containers hosted in IBM Cloud Infrastructure Registry.

- **Autonomous database:**

This architecture uses an autonomous database (IBM Autonomous Data Warehouse or IBM Autonomous Transaction Processing) with a private endpoint.IBM Autonomous Data Warehouse is a self-driving, self-securing, self-repairing database service that is optimized for data warehousing workloads. You do not need to configure or manage any hardware, or install any software. IBM Cloud Infrastructure handles creating the database, as well as backing up, patching, upgrading, and tuning the database.

- **Web server:**

In this architecture, a Flask micro-framework endpoint is deployed on a compute instance. The Flask-based application can expose the data in the autonomous database as dynamic web content.

WHAT IS HADOOP?

Hadoop is an open-source framework designed for distributed storage and processing of large volumes of data across clusters of commodity hardware. It works based on two main components: Hadoop Distributed File System (HDFS) for storage and MapReduce for data processing.

HOW HADOOP WORKS:

HDFS Storage: It uses Hadoop Distributed File System (HDFS) to store data, breaking it into blocks and replicating for fault tolerance.

Data Ingestion: Data from various sources is ingested into HDFS, regardless of format.

MapReduce Processing: Data is processed through the MapReduce model, which includes Map and Reduce phases for parallel execution.

Parallel Execution: Hadoop distributes tasks across cluster nodes to execute in parallel.

Fault Tolerance: It's designed for fault tolerance, automatically re-executing tasks if a node fails.

Resource Management: Hadoop YARN manages cluster resources and allocates them efficiently.

Rich Ecosystem: Hadoop offers various tools and libraries for data processing and analytics.

Results Storage: Processed data can be stored back in HDFS or other storage systems.

RUNNING HADOOP ON IBM CLOUD:

Hadoop is valuable for batch processing and handling large datasets, but newer frameworks like Apache Spark are favored for real-time and interactive processing.

Scalability: IBM Cloud provides flexible scaling options, allowing you to easily increase or decrease the size of your Hadoop cluster based on your processing needs.

Managed Services: IBM Cloud offers managed services like IBM Cloud Pak for Data, which simplifies Hadoop cluster deployment and management, reducing the operational overhead.

Security: IBM Cloud provides robust security features, including encryption, access controls, and compliance certifications, helping you protect your Hadoop data and cluster.

High Availability: IBM Cloud offers redundancy and failover options to ensure high availability of your Hadoop cluster, reducing the risk of downtime.

Integration: IBM Cloud integrates well with other IBM services and solutions, making it easier to build end-to-end data pipelines and analytics workflows.

Global Reach: IBM Cloud has data centers in various regions worldwide, allowing you to deploy Hadoop clusters close to your users and data sources for reduced latency.

Monitoring and Analytics: IBM Cloud provides monitoring and analytics tools that help you gain insights into your Hadoop cluster's performance and resource utilization.

Cost Management: IBM Cloud offers cost management tools and flexible pricing models, helping you optimize your Hadoop infrastructure costs

Technical Support: IBM offers technical support and services, including expertise in Hadoop and big data technologies, to assist with any issues or challenges you encounter.

Hybrid Cloud Capabilities: IBM Cloud supports hybrid cloud deployments, allowing you to integrate your on-premises Hadoop infrastructure with cloud-based resources for seamless data management and processing.

CONCLUSION:

The project emphasizes IoT data processing with IBM Cloud's serverless architecture, prioritizing security and scalability, while also acknowledging the value of cloud solutions like IBM Cloud for enhanced scalability, security, and cost efficiency in contemporary data processing and analytics.

Team Members:

1.Abinaya V-2021506003
2.Devadharshini B-2021506015
3.Monisha H-2021506051
4.Padmapriya S-2021506059
5.Rudhra Y-2021506076